5.5

Cairo University
Faculty of Computers and Artificial Intelligence

Midterm Exam

Department: CS
Course Name: Machine Learning
Course Code: CS467 / SCS467
Instructor(s): Dr. Hanaa Bayomi

Date: 22/11/2022
Duration: 1 hour
Total Marks: 20

[redacted]

تعليمات هامة

- جهازة التليفون المحمول مقفوحا داخل لجنة الإمتحان يعتبر حالة غش تستوجب العقاب وإذا كان ضرورى الدخول بالمحمول فيوضع مغلق فى الحقائب.
- لا يسمح بدخول سماعة الأذن أو البلوتوث.
- لايسمح بدخول أي كتب او ملازم أو أوراق داخل اللجنة والمخالفة تعتبر حالة غش.

Question 1 *[11 marks]*

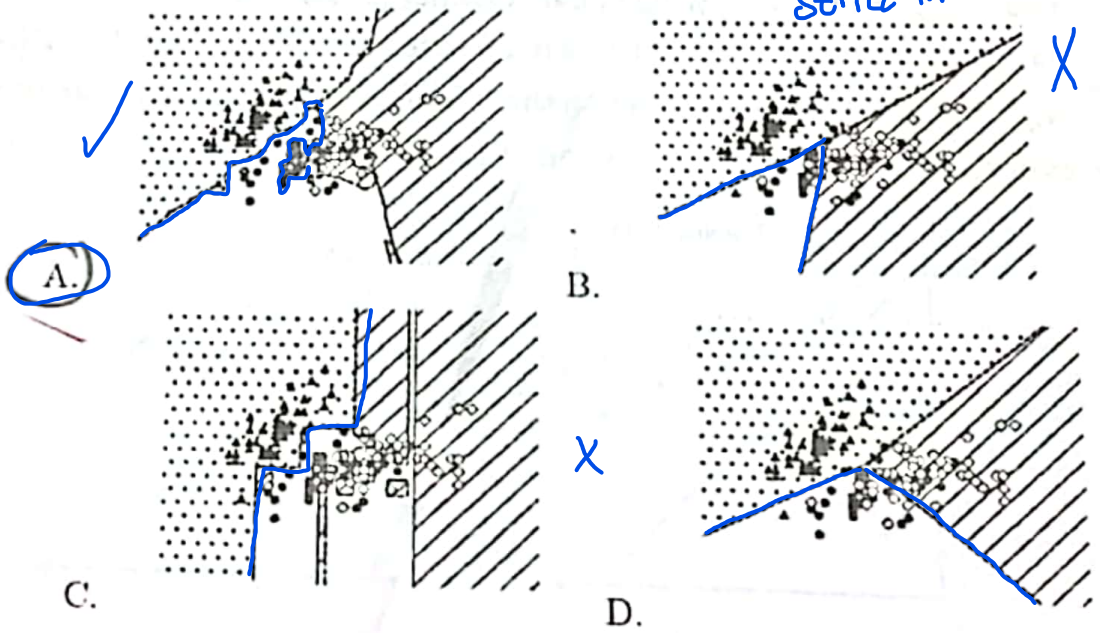
- Answer the following Questions:

1. We need re-estimate probabilities (smoothing) in Naïve Bayes classifier. [1 mark]

yes, because feature are independend

Yes, to avoid zero conditional probability problem, which happens when no example contains one of the feature values provided in the input (Xi).

2. Which of the following decision boundaries is most likely to be generated by a k-NN? Why? [2 marks]



Because it is interested in closed and works on live

The decision boundary generated by k-NN is typically non-linear and depends on the local structure of the data points.

3. For the following dataset, calculate the error rate for each KNN classifier. [2 marks]

(b) How many samples are incorrectly classified on the test data for Class 1 (+) and Class 2 (*)? [1 mark]

* 2

\+ 1 + Zero

Question 2 *[9 marks]*

You are stranded on a deserted island. Mushrooms of various types grow wildly all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous and others as not (determined by your former companions' trial and error). You are the only one remaining on the island. You have the following data to consider.

	IsHeavy	IsSmelly	IsSpotty	IsSmooth	IsPoisonous
A	0	0	1	0	0
B	1	1	0	1	0
C	1	0	0	1	1
D	0	1	1	0	1
E	0	0	0	1	1
F	0	0	1	1	1
G	1	1	0	0	1
H	1	1	0	0	1
U	1	1	1	1	?
V	0	1	0	1	?
W	1	1	0	0	?

You know whether or not mushrooms A through H are poisonous, but you do not know about U through W. For the first couple of questions, consider only mushrooms A through H.

a) **What is the entropy of IsPoisonous? [1.5 mark]**

$= -P_0 \log_2 P_0 - P_1 \log_2 P_1$

$P_0 = \frac{3}{8}$, $P_1 = \frac{5}{8}$

entropy $= -\frac{3}{8}\log\left(\frac{3}{8}\right) - \frac{5}{8}\log\left(\frac{5}{8}\right)$

$= 0.954$

b) **Which attributes should you choose as the root of decision tree? [3 mark]**

$H(root) = 0.954$

Heavy

=0 → 3 yes, 2 no (2) (0)

$E(Hv=0) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$

$= 0.97$

=1 → 2 yes, 1 no

$E(Hv=1) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}$

$= 0.92$

$G(root, heavy) = 0.95 - \frac{5}{8}(0.97) - \frac{3}{8}(0.92) = 3.2 \times 10^{-3}$

Smelly

=0 → 2 no, 3 yes

$E(Sm=0) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.971$

=1 → 1 no, 2 yes

$E(Sm=1) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918$

$G(r, Sm) = 0.954 - \frac{5}{8}(0.971) - \frac{3}{8}(0.918) = 3.2 \times 10^{-3}$

Sp

=0 → 2 no, 3 yes → Same as previous

=1 → 1 no, 2 yes → Same

$G(r, Sp) = 3.2 \times 10^{-3}$

Smath

=0 → 2 no, 2 yes

$E(Sm=1) = 1$

=1 → 1 no, 3 yes

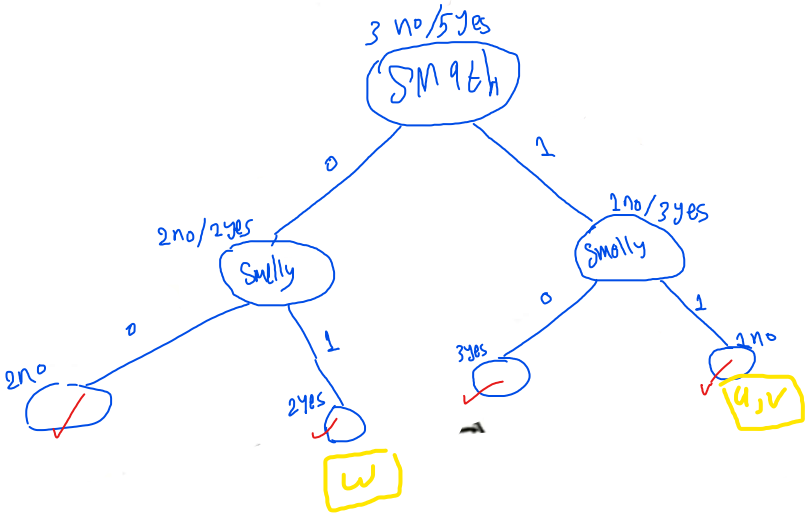
$E(Sm=1) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0.811$

$G(r, Smath) = 0.954 - \frac{4}{8}(1) - \frac{4}{8}(0.811)$

$= 0.0488$

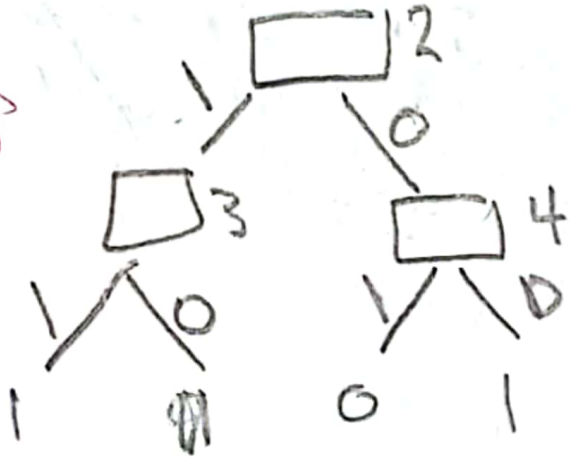
✳✳ Highest Gain

c) *Build decision tree to classify mushrooms as poisonous or not?* **[3 marks]**



d) **Classify mushrooms U,V and W using this decision tree as poisonous or not?**
[1.5 marks]

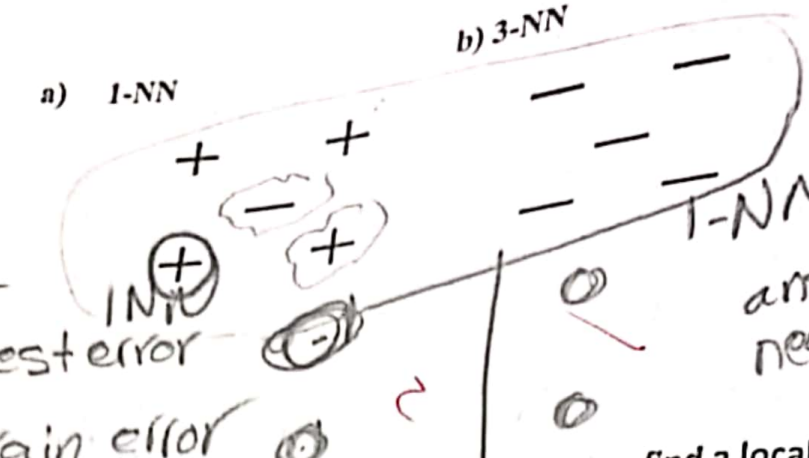
U, V → No
W → yes



Good
Dr. Hanaa B

a) 1-NN

b) 3-NN



1-NN answer negative | 3NN answer negative

1NN test error 0

train error 0

4. (True or False): Gradient descent is guaranteed to find a local minimum [1 mark]

False ~~True~~

True, if α is correctly chosen

5. Suppose you run gradient descent for linear regression for 100 iterations with a learning rate 0.01. You observe that the training loss (sum of squares error) is increasing after every iteration. How would you explain this observation? And further, what changes would you make to the set up for the gradient descent to converge to a solution? [2 Marks]

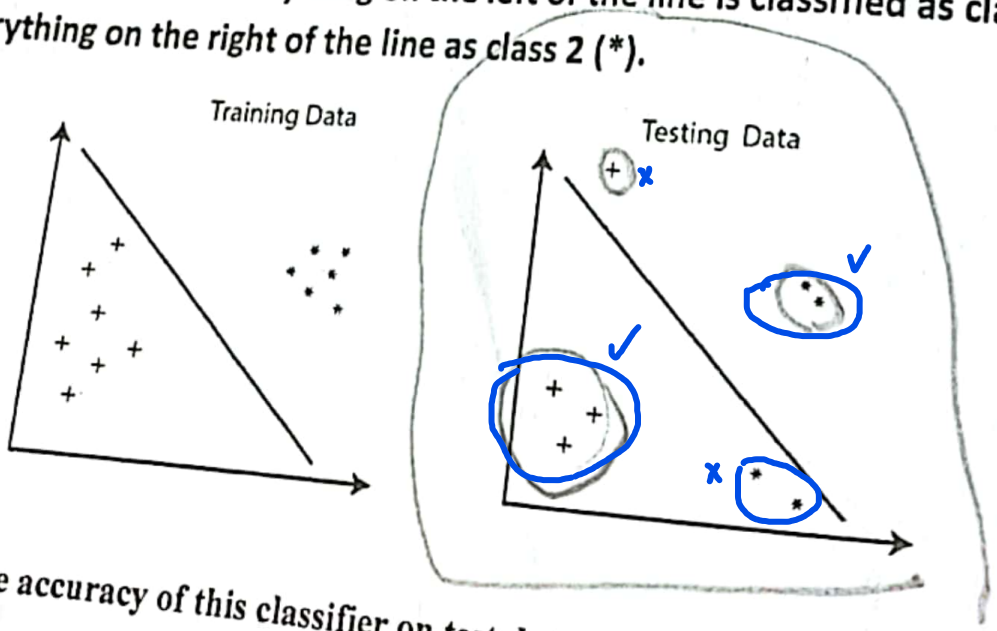
The learning rate is too big that it caused divergence (moving further from the minimum). I'd update the learning rate to a smaller number and check if the training loss is decreasing with iterations, else I'd decrease more until it converges.

learn rating is small, need to increase

data is not normalization, Make data normalize

Scale by mean

6. You are given the following linear classifier (shown on both training and testing data), with data belonging to class 1 (represented with +) or class 2 (represented with *). Note that everything on the left of the line is classified as class 1 (+), and everything on the right of the line as class 2 (*).



What is the accuracy of this classifier on test data for Class 1 (+) and Class 2 (*)?[2 marks]

a) $\frac{TP+TN}{\text{total}} = \frac{\text{correct } 2 \leftarrow 5}{5 \text{ total} \leftarrow 8}$

b) 3